HANDS ON SCRAPING DATA DENGAN KASUS ANALISIS SENTIMEN PADA KEBIJAKAN PEMERINTAH

Wahyu Calvin Frans Mariel – BPS
Muhammad Ghozy Al Haqqoni – BPS

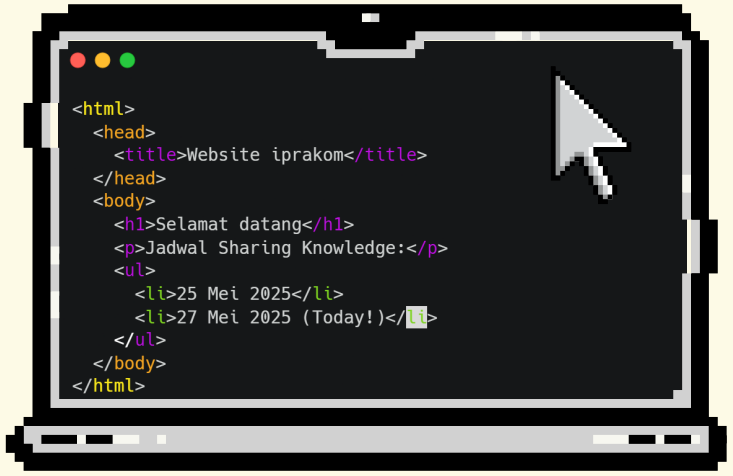
Rabu, 28 Mei 2025

WEB SCRAPING 101

Muhammad Ghozy Al Haqqoni

Rabu, 28 Mei 2025

Web Scraping?



- ✓ Teknik **mengumpulkan data** yang jumlahnya bisa sedikit hingga sangat besar.
- ✓ Berisi berbagai informasi yang disediakan oleh suatu *website*.
- ✓ Biasanya dilakukan setelah proses pengumpulan URL dari *web crawling*.

Web Crawling?


```
<html>
  <head>
    <title>Website iprakom</title>
  </head>
  <body>
    <h1>Selamat datang</h1>
    <p>Jadwal Sharing Knowledge:</p>
    <ul>
      <li>25 Mei 2025</li>
      <li>27 Mei 2025 (Today!)</li>
    </ul>
  </body>
</html>
```

- ✓ Suatu proses **menjelajahi halaman** suatu *website* yang biasanya memprogram untuk *indexing* seluruh web.
- ✓ Bisa digunakan untuk meningkatkan kualitas SEO (*search engine optimization)* suatu web. Seperti Googlebot, Skyscanner, webarchive.

Objective

1. ***Netiquette***: Semua *website* boleh di-*scraping*?
2. Mengenal lebih dalam: ***Website***
3. ***Tools*** untuk *web-scraping*
4. ***Hands-on***: *Real world cases*

Netiquette: Web Scraping Ethics

Penting!

Konsiderasi memilih *website* yang dapat dilakukan *web scraping*

---

1. *Term and Conditions* Websites
2. Robots.txt (*easier way*)
3. *Cloudflare* atau *anti-bot software* lainnya

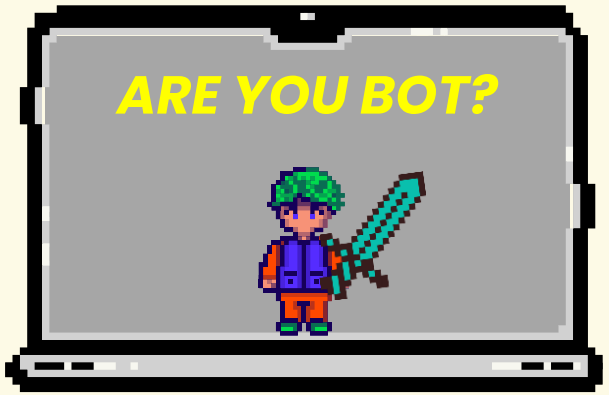
Netiquette: Robots.txt

File robots.txt biasanya digunakan oleh pemilik *website* untuk mengendalikan akses dan memberikan instruksi pada bot mesin pencari tentang ***rules* apa yang diperbolehkan dan dilarang** karena alasan privasi atau keamanan.

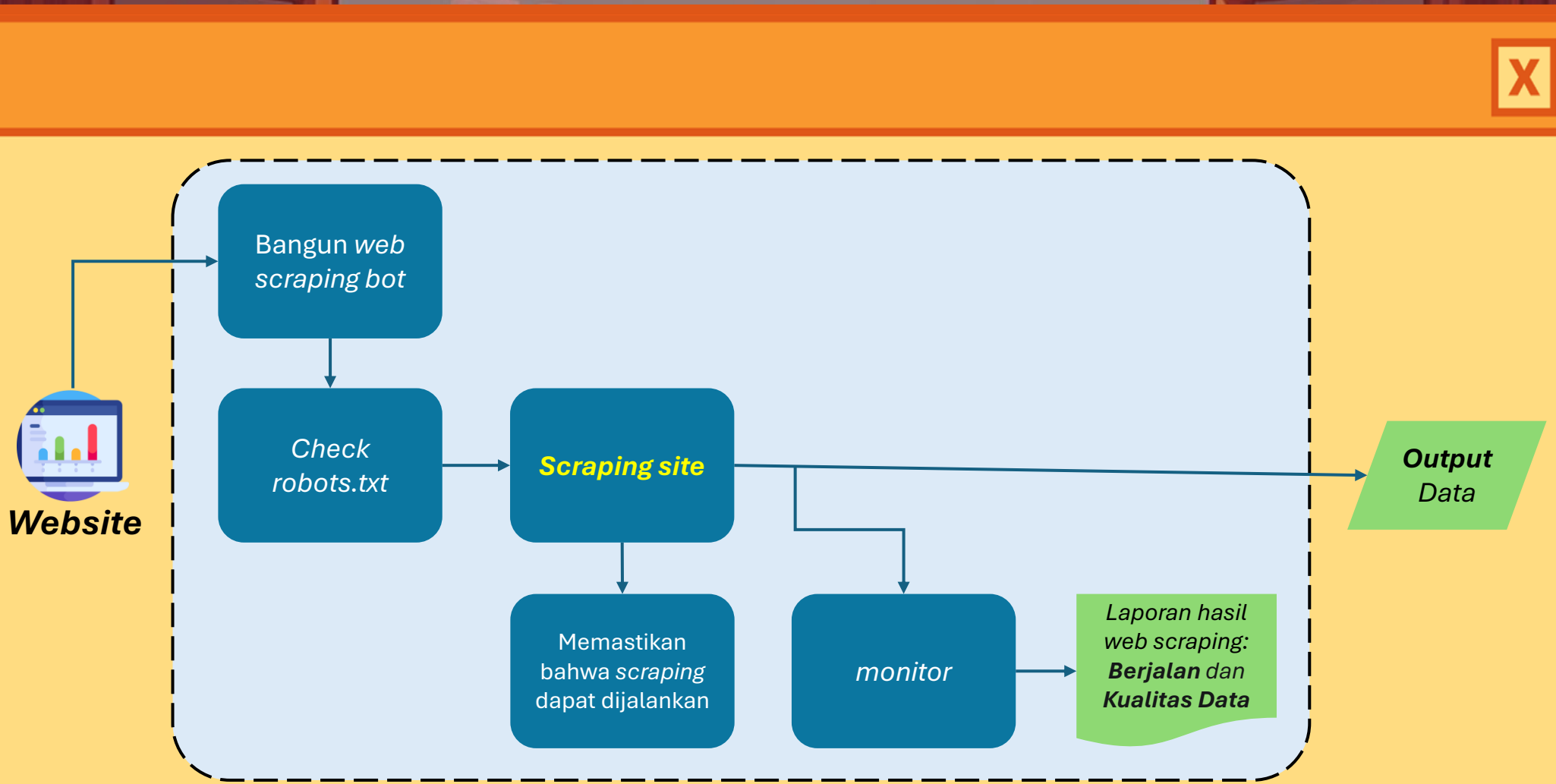
Bisa cek dari https://(websiteyangdituju)/robots.txt

kompas.com/robots.txt

```
User-Agent: *
Disallow: /search/?q=
Disallow: /komentar/*
Disallow: /copy/*
Disallow: *?jxrecoid=*
Disallow: *?utm_source=*
Disallow: *?source=*
```

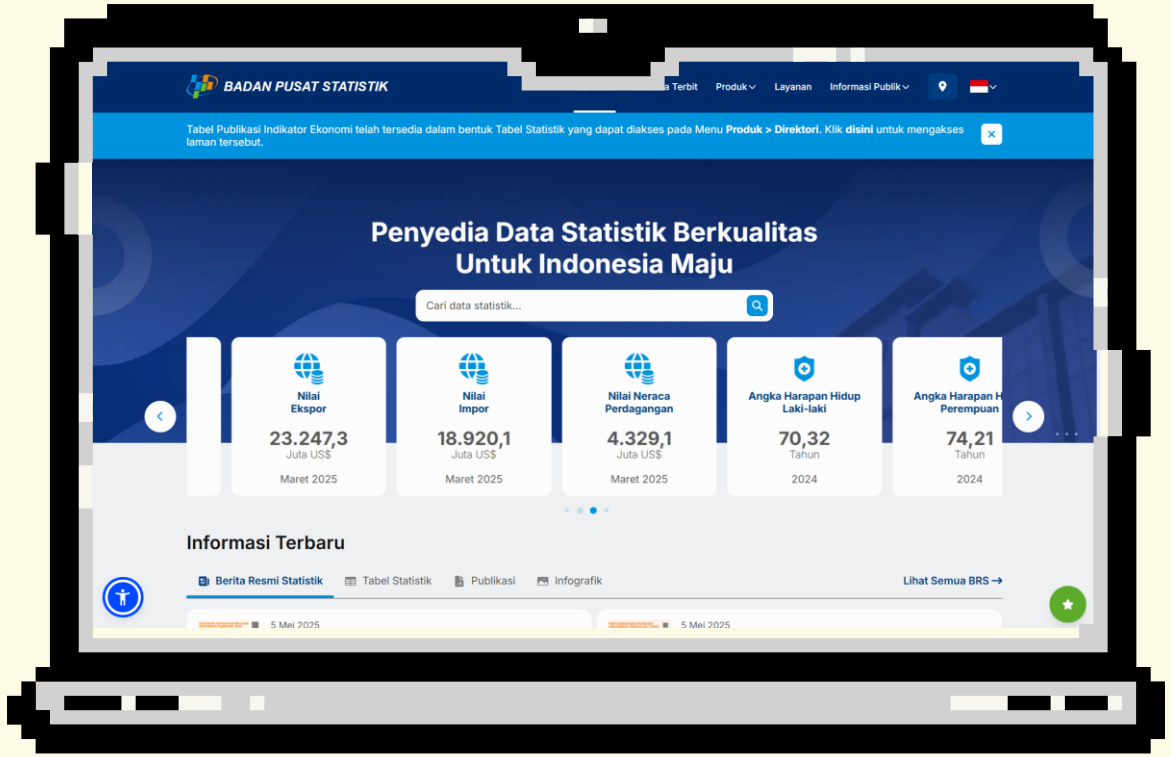


Guidance to Web Scraping



Website

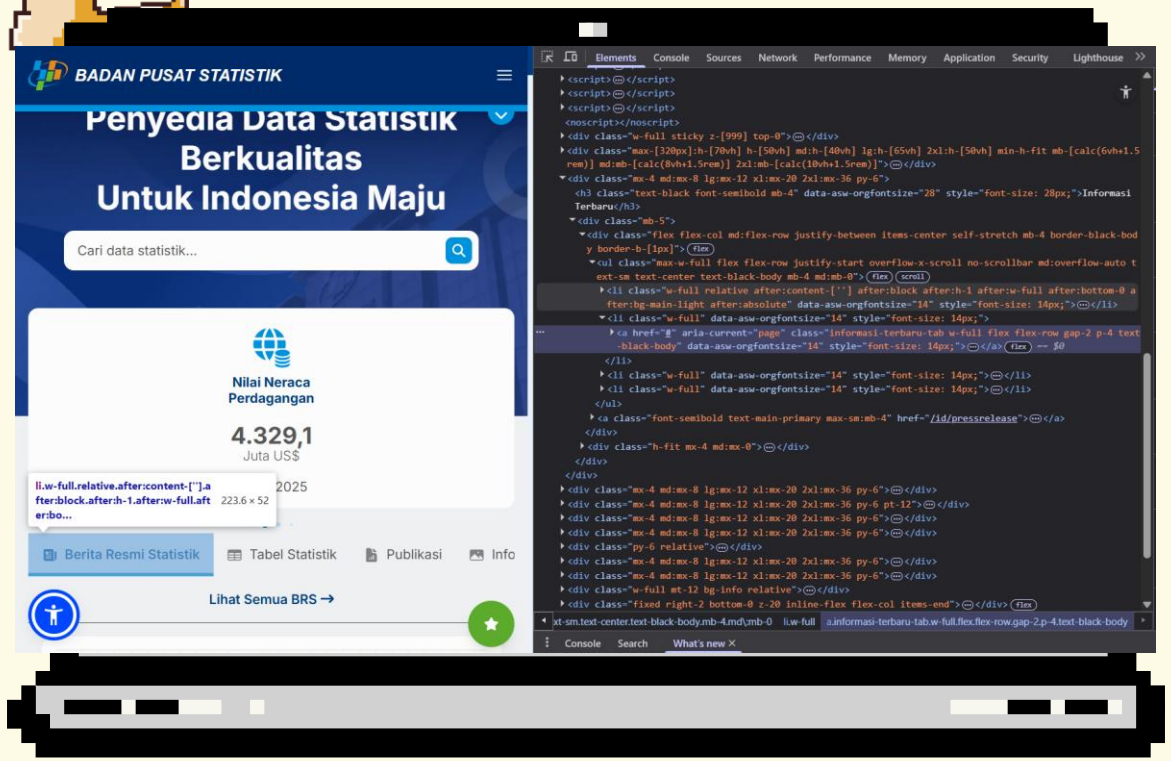
Source: *https://www.bps.go.id*



Kumpulan *web pages* yang bisa diakses lewat alamat tertentu

Website

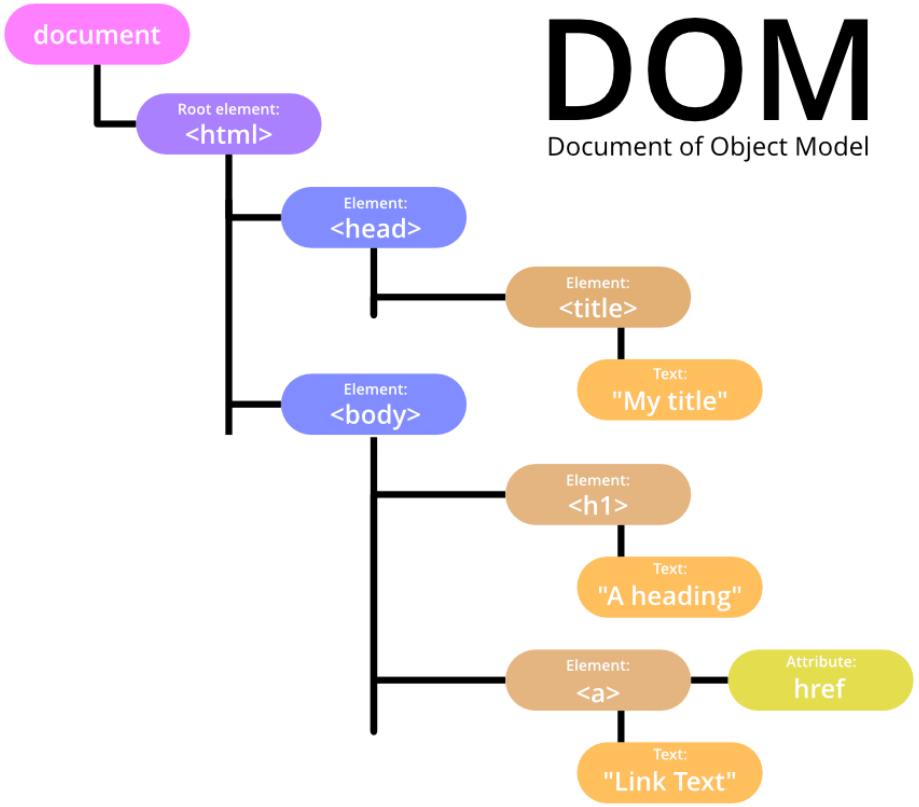
Source: *https://www.bps.go.id*



Belakang layar

Dibangun dengan *programming language* dan di-*load* melalui HTML

HyperText Markup Language



Source: *https://www.geeksforgeeks.org/what-is-document-object-in-java-dom/*

Struktur dari suatu *website*

- ✓ Membangun dan menghubungkan suatu web pages satu dan lainnya
- ✓ ,Menampilkan berbagai informasi: *head* and *body*

HTML: Tags

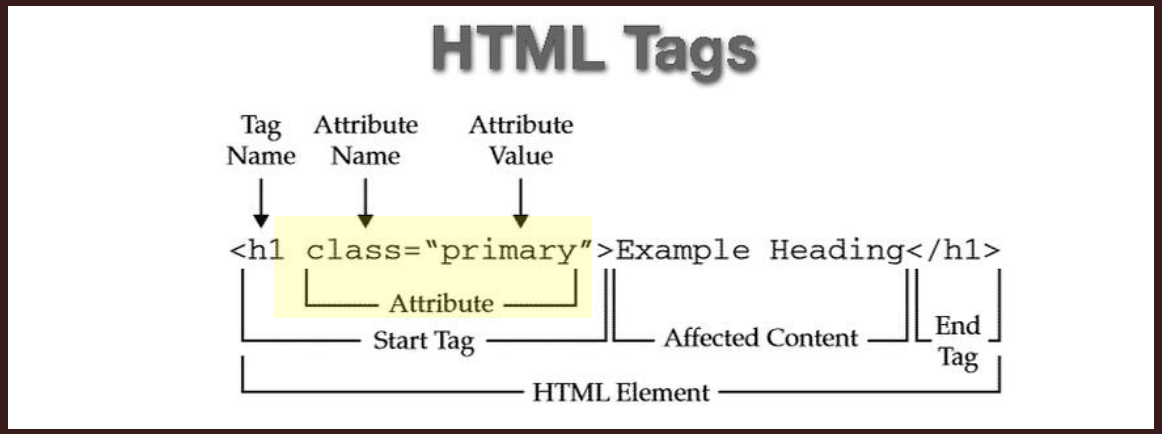
Tag	Description
<html> ... </html>	Declares the Web page to be written in HTML
<head> ... </head>	Delimits the page's head
<title> ... </title>	Defines the title (not displayed on the page)
<body> ... </body>	Delimits the page's body
<h *n*> ... </h*n*>	Delimits a level *n* heading
...	Set ... in boldface
...	Set ... in italics
...	Center ... on the page horizontally
...	Brackets an unordered (bulleted) list
...	Brackets a numbered list
...	Brackets an item in an ordered or numbered list
 	Forces a line break here
<p>	Starts a paragraph
<hr>	Inserts a horizontal rule
<img src="...">	Displays an image here
 ... 	Defines a hyperlink

Source: *https://medium.com/@nwachukwufavourc/understanding-html-tags-and-elements-795402e3eb47*

Simbol untuk menandai suatu elemen di HTML

- ✓ Berisi instruksi kepada *browser* bagaimana suatu objek bisa ditampilkan
- ✓ objek bisa berupa text, video, audio, dan gambar

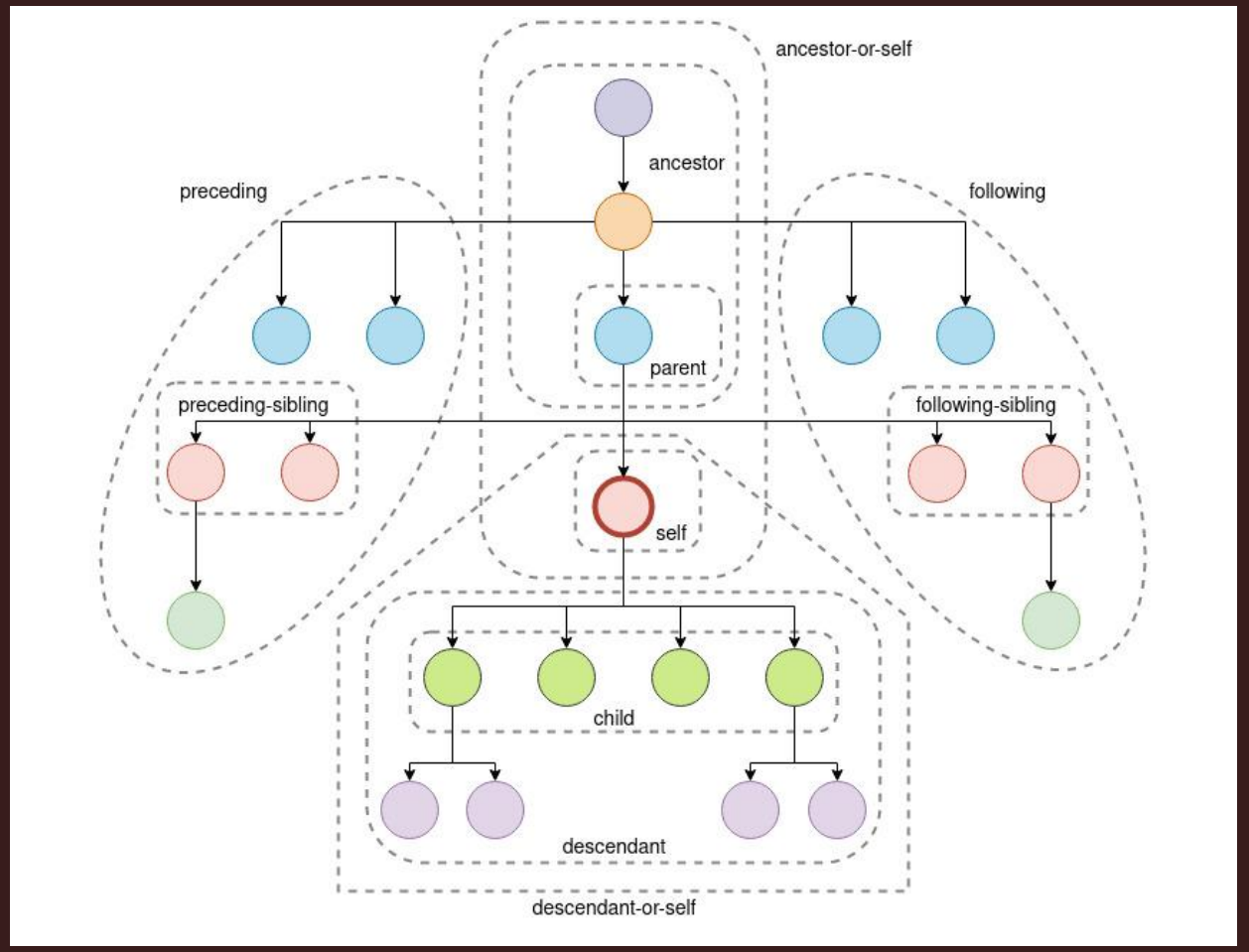
Tags Attributes



Source: *https://medium.com/@nwachukwufavourc/understanding-html-tags-and-elements-795402e3eb47*

Komponen untuk mengidentifikasi sebuah *tag*

Tags Axes



Source: *https://jrebecchi.github.io/xpath-helper/xpath-axes.html#parent-axis/*

Identifikasi *tags* berdasarkan hubungan antar elemen

- ✓ Menavigasi hierarki atas/bawah *tags*
- ✓ *Parent-to-sibling, Parent-to-child, dsb.*

```
<html>
  <head>
    <title>Website iprakom</title>
  </head>
  <body>
    <h1>Selamat datang</h1>
    <p>Jadwal Sharing Knowledge:</p>
    <ul>
      <li>25 Mei 2025</li>
      <li>27 Mei 2025 (Today!)</li>
    </ul>
  </body>
</html>
```

Contoh:

li *: child to ul*

p *: sibling to h1 and ul*

The Way to Select Tag(s)

CSS
SELECTORS

XPATH

The Way to Select Tag(s)

Fitur	**CSS** SELECTORS	**XPATH**
Sintaks	Mirip CSS	Lebih kompleks
Arah navigasi	Dari atas ke bawah saja	Bisa naik-turun (ke *parent*, ke *child*, ke *sibling*)
Kompatibilitas	BeautifulSoup	lxml
Baik digunakan saat	Struktur HTML sederhana	Struktur kompleks atau banyak kondisi

The Way to Select Tag(s)

PROMPT:

1. ambil <h2 class="title"> dalam
2. semua <a> dengan href diawali https://
3. elemen <li> di dalam <ul class="menu"> dalam
4. elemen yang punya anak <p class="desc">

CSS
SELECTORS

```
div.article h2.title
a[href^="https://"]
div#header > ul.menu > li
```

XPATH

```
//div[@class="article"]//h2[@class="title"]
//a[starts-with(@href, "https://")]
//div[@id="header"]/ul[@class="menu"]/li
//p[@class="desc"]/parent::div
```

Web Scraping Tools

- ✓ MANUALLY
- ✓ BANTUAN PIHAK KETIGA
- ✓ GOOGLE SHEETS
- ✓ JAVASCRIPT
- ✓ R
- ✓ PYTHON
- ✓ AND SO ON...

Python Libraries

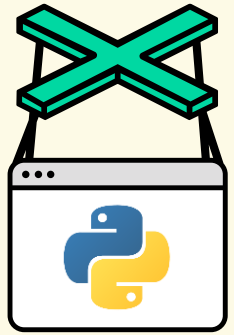
Requests

Beautiful Soup

Se

SCRAPLING

Scrapy



Type of Web Scraping

HTML Extraction

Mengumpulkan data dengan mengekstraksi informasi yang ada dalam struktur HTML *web page*

API Extraction

Mengumpulkan data melalui API (*Application Programming Interface*) yang disediakan suatu *website.* Biasanya *output* dalam JSON

Bisa dilakukan pengecekan pada *tab network* di *developer tools browser*:
Windows: Ctrl + Shift + I
MacOS: Command + Option + I (⌘+⌥+I)


ANY QUESTION
SO FAR?

LET’S RUN A QUICK SIMULATION

Kalau mendengar **kebijakan pemerintah**, yang terpikirkan oleh teman-teman, kita bisa **memperoleh informasi** terkait kebijakan itu **dari mana** saja?

SOSIAL MEDIA

BERITA

YOUTUBE

WHATSAPP CHANNEL

DENGER LANGSUNG


LET'S DO
SOME PRACTICE

Rabu, 28 Mei 2025

THANKYOU

contact us:

Wahyu Calvin Frans Mariel: linkedin.com/in/wahyu-calvin
Muhammad Ghozy Al Haqqoni : linkedin.com/in/mghozyah

Lampiran:
The Way to Select Tag(s)

CSS
SELECTORS

Tujuan	Contoh CSS Selector	Penjelasan
Pilih elemen berdasarkan tag	div	Semua
Berdasarkan kelas	.container	Elemen dengan class="container"
Berdasarkan id	#header	Elemen dengan id="header"
Tag + kelas	div.article	dengan class="article"
Tag + id	div#main	dengan id="main"
Atribut tertentu	a[href="https://"]	<a> dengan href persis "https://"
Atribut yang dimulai dengan	a[href^="https"]	href mulai dengan "https"
Child langsung	div > p	<p> yang langsung anak
Semua descendant	div p	Semua <p> di dalam
Sibling langsung	h2 + p	<p> langsung setelah <h2>
Semua sibling	h2 ~ p	Semua <p> setelah <h2> pada level yang sama

Lampiran:
The Way to Select Tag(s)

XPATH

Tujuan	Contoh XPath	Penjelasan
Semua elemen tag tertentu	//div	Semua di dokumen
Berdasarkan atribut kelas	//div[@class="container"]	dengan class="container"
Berdasarkan atribut id	//div[@id="header"]	dengan id="header"
Tag + atribut tertentu	//a[@href="https://"]	<a> dengan href persis "https://"
Atribut yang dimulai dengan	//a[starts-with(@href, "https")]	href mulai dengan "https"
Child langsung	/div/p	<p> yang langsung anak
Semua descendant	//div//p	Semua <p> di dalam
Naik ke parent	//p[@class="desc"]/parent::div	Parent dari <p class="desc">
Pilih berdasarkan posisi	//ul/li[3]	ketiga dalam
Pilih berdasarkan teks	//a[contains(text(), "Berita")]	<a> yang mengandung kata "Berita"